Nonlinear dynamics of river runoff elucidated by horizontal visibility graphs

Holger Lange, Sebastian Sippel, and Osvaldo A. Rosso

Citation: Chaos **28**, 075520 (2018); doi: 10.1063/1.5026491
View online: https://doi.org/10.1063/1.5026491
View Table of Contents: http://aip.scitation.org/toc/cha/28/7
Published by the American Institute of Physics

Chaos
An Interdisciplinary Journal of Nonlinear Science
Fast Track Your Research. Submit Today!

Nonlinear dynamics of river runoff elucidated by horizontal visibility graphs

Holger Lange,[1,a)] Sebastian Sippel,[1] and Osvaldo A. Rosso[2,3,4]

[1] *Norwegian Institute of Bioeconomy Research, Postboks 115, N-1433 Ås, Norway*
[2] *Departamento de Informática en Salud, Hospital Italiano de Buenos Aires and CONICET, C1199ABB Ciudad Autónoma de Buenos Aires, Argentina*
[3] *Instituto de Física, Universidade Federal de Alagoas, Av. Lourival Melo Mota, s/n, 57072-970 Maceió, Alagoas, Brazil*
[4] *Complex Systems Group, Facultad de Ingeníera y Ciencias Aplicadas, Universidad de los Andes, Las Condes, 12455 Santiago, Chile*

(Received 20 February 2018; accepted 19 June 2018; published online 31 July 2018)

Horizontal Visibility Graphs (HVGs) are a recently developed method to construct networks from time series. The values of the time series are considered as the nodes of the network and are linked to each other if there is no larger value between them, such as they can "see" each other. The network properties reflect the nonlinear dynamics of the time series. For some classes of stochastic processes and for periodic time series, analytical results can be obtained for network-derived quantities such as the degree distribution, the local clustering coefficient distribution, the mean path length, and others. HVGs have the potential to discern between deterministic-chaotic and correlated-stochastic time series. Here, we investigate the sensitivity of the HVG methodology to properties and pre-processing of real-world data, i.e., time series length, the presence of ties, and deseasonalization, using a set of around 150 runoff time series from managed rivers at daily resolution from Brazil with an average length of 65 years. We show that an application of HVGs on real-world time series requires a careful consideration of data pre-processing steps and analysis methodology before robust results and interpretations can be obtained. For example, one recent analysis of the degree distribution of runoff records reported pronounced sub-exponential "long-tailed" behavior of North American rivers, whereas another study of South American rivers showed hyper-exponential "short-tailed" behavior resembling correlated noise. We demonstrate, using the dataset of Brazilian rivers, that these apparently contradictory results can be reconciled by minor differences in data-preprocessing (here: small differences in subtracting the seasonal cycle). Hence, data-preprocessing that is conventional in hydrology ("deseasonalization") changes long-term correlations and the overall runoff dynamics substantially, and we present empirical consequences and extensive simulations to investigate these issues from a HVG methodological perspective. After carefully accounting for these methodological aspects, the HVG analysis reveals that the river runoff dataset shows indeed complex behavior that appears to stem from a superposition of short-term correlated noise and "long-tailed behaviour," i.e., highly connected nodes. Moreover, the construction of a dam along a river tends to increase short-term correlations in runoff series. In summary, the present study illustrates the (often substantial) effects of methodological and data-preprocessing choices for the interpretation of river runoff dynamics in the HVG framework and its general applicability for real-world time series. *Published by AIP Publishing.* https://doi.org/10.1063/1.5026491

We study the dynamics of water flow, given as time series of river runoff from long-term measurement stations (up to 85 years of daily data) in Brazil. The time series are analyzed using "Horizontal Visibility Graphs." In this method, time series are represented as a network: each value of the time series is a node of the network, and two nodes are linked to each other if they can "see" each other in the horizontal direction (no higher values are in between them), i.e., analogous to horizontal visibility in a landscape. Properties of the network provide insight into the temporal structure of the river runoff; in particular, it can be determined to which extent river runoff resembles certain types of random processes. We demonstrate that the analysis has to be carried out with great care in order to avoid misinterpretations and wrong conclusions. In particular, we show the consequences of the presence of identical values in the time series, of different versions of taking out the seasonal trend, and of the finite length of the series. For the latter, we use computer-generated data from random processes, where analytical results for infinite length are known. If thoroughly applied, the Horizontal Visibility Graphs are tools for the analysis of time series providing insights into the dynamics and a presentation of their behavior not easily obtained otherwise.

I. INTRODUCTION

Complex networks constructed from time series of (Earth) observations, univariate or multivariate, have become increasingly popular in recent years.[1–4] The network approach provides insight into dynamics of systems in a different way

[a)] Author to whom correspondence should be addressed: holger.lange@nibio.no than other methods; they may help for the classification of systems according to their network structure or topological properties,[5] but also ubiquitous features like the "small-world"[6] or the "scale-free"[7] properties have been observed.

In this contribution, we focus on a method to generate a network from a univariate time series which is among the conceptually most simple analysis techniques: the Horizontal Visibility Graphs (HVGs).[8] Every data point (observation) of a time series x is considered as a node of the network to be formed; two nodes at observation times t_i and a later t_j are connected by a link iff none of the values in between them is larger than either of the two:

$$x_k < \inf(x_i, x_j) \quad \forall k : \; i < k < j, \tag{1}$$

which implies that the two time series values can "see" each other when looking horizontally. Moreover, there is an option to consider the links between nodes as arrows (directed HVGs) or as lines (undirected HVGs). The consideration of directed graphs opens for the possibility to analyze differences in the time direction (irreversibility).[9]

The set of nodes and links constitutes the graph, which can be visualized or formally expressed as adjacency matrix A. The simplest choice for A is binary: the entry A_{ij} is 1 if the two nodes i and j are linked, and 0 otherwise. The alternative is a weighted adjacency matrix where the matrix elements are related, e.g., to the difference in time series values of the nodes linked together,[10] or proportional to the temporal distance between the linked values. Either way, a characteristic of the adjacency matrix is that it is sparse (many 0s) for typical time series, which is convenient for long time series involving big matrices.

We investigate time series of river runoff rates at daily resolution. The runoff at a given location is the result of interactions between precipitation, air temperature and other meteorological variables, vegetation, soils, and the geophysical system (catchment) upstream the measurement gauge. Already since the seminal work of Hurst,[11] it has been known that a typical characteristic of runoff time series is their persistence, or long-range dependence, indicated, e.g., by autocorrelation functions decaying slower than exponential as a function of the temporal lag. In addition, runoff data contain periodicities, foremost the annual cycle, but also multiyear structures and long-term trends, and are not the least influenced by human management, e.g., channel regulation or water power generation. In addition, information on runoff rates is often economically relevant, as knowledge about the magnitude of extreme events (droughts and floods) and the response time to rainfall events guide the construction of infrastructure for protection and utilization of the water resources. Thus, time series from river runoff comprise a relevant domain for data analysis, ecosystem research, risk analysis, and also climate change research.[12,13]

The paper is organized as follows:

In the first part of the analysis, we investigate in detail methodological and data pre-processing choices that are crucial for an application of the HVG framework to real-world time series. For this purpose, we analyze (1) artificially generated time series of varying length and autocorrelation structure and (2) the Brazilian river runoff dataset as an illustrative example. For the artificial time series, we focus on

(1) Sample size effects on HVGs itself and the estimation method of HVG-based summary statistics (i.e., the "lambda" parameter, explained in Sec. II), which are an issue for short time series;
(2) The effect of different deseasonalization procedures on the tail behavior of the degree distributions;
(3) The effect of ties in the time series.

In the second part of the analysis, we apply the HVG framework to the Brazilian runoff dataset with a focus on (1) the short- vs. long-tailed behavior of these time series and (2) the effect of dam construction along these rivers on their corresponding HVG-based degree distributions. A flow chart that illustrates the structure of the paper and the different data, analyses, and corresponding results is shown in Fig. 1.

II. MATERIALS AND METHODS

A. Dataset

We analyze river runoff time series (unit m^3/s) from a total of 146 stations, all located in Brazil. They were obtained from the Brazilian federal institution that controls the electric power production, the Operador Nacional do Sistema Elétrico (ONS). [The data were obtained from http://www.ons.org.br/operacao/vazoes_naturais.aspx (webpage not available as of March 2018).] Hydropower is the

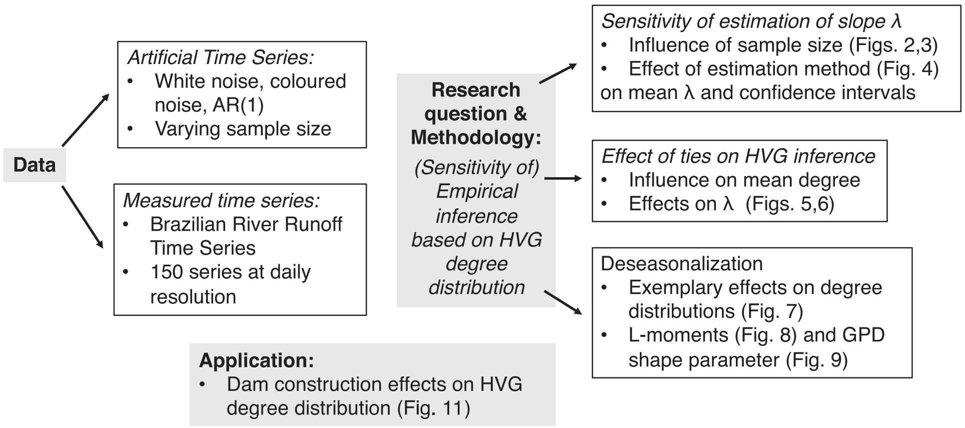
Data
Artificial Time Series:
• White noise, coloured noise, AR(1)
• Varying sample size
Measured time series:
• Brazilian River Runoff Time Series
• 150 series at daily resolution
Research question & Methodology:
(Sensitivity of) Empirical inference based on HVG degree distribution
Sensitivity of estimation of slope λ
• Influence of sample size (Figs. 2,3)
• Effect of estimation method (Fig. 4)
on mean λ and confidence intervals
Effect of ties on HVG inference
• Influence on mean degree
• Effects on λ (Figs. 5,6)
Deseasonalization
• Exemplary effects on degree distributions (Fig. 7)
• L-moments (Fig. 8) and GPD shape parameter (Fig. 9)
Application:
• Dam construction effects on HVG degree distribution (Fig. 11)

FIG. 1. Flow chart of different datasets used, analyses, and corresponding results in this paper.

single most important source of electricity in the country, contributing around 75% to the total production, with Brazil being the third largest hydropower producer world-wide.

The first entries of the data record date back to 1931; the median length of the daily time series is 27 000 values (74 years of data), the measurements extend towards the end of 2014. The catchment sizes vary between 250 km^2 and just below 1×10^6 km^2. A closely related, slightly shorter dataset has also been analyzed in Ref. 14.

Each of the 146 time series originates from a river where a dam was constructed; however, the majority of the dams were built in more recent years: 87% of the gauges were impacted by dam construction after 1960, and 61 gauges (42%) even only after 2000. Thus, the impact of direct management of water flows, seen on the scale of the whole record for each gauge, varies significantly. Since the year of dam construction is known for every gauge, we can quantify the anthropogenic impact by splitting the time series in the periods before and after the construction, whenever this is meaningful, and perform the analysis separately for the two parts.

B. Seasonality and its approximate removal

A dominant feature of runoff data is their seasonality, induced by seasonal patterns in precipitation or (to a lesser extent in Brazil) the annual cycle of temperatures. The amplitude of the seasonal cycle (or the fraction of variance explained by the annual signal) differs a lot between the gauges; in addition, it is to be expected that the presence of seasonality could have a profound influence on the results of a Horizontal Visibility Graph analysis. Following common practice in hydrology, we therefore construct deseasonalized datasets from the original one using the following recipe:

Let $x_{y,i}$ be a value of the time series, obtained in year y and on day of the year i. We calculate the mean value for each given day of the year from all the years, μ_i, and the corresponding standard deviation σ_i, $i = 1, \ldots, 365$ (for leap years, February 29 has been removed from the analyses for simplicity). Then, the original values of the time series are transformed according to

$$x_{y,i}^{deseas} = \frac{x_{y,i} - \mu_i}{\sigma_i}. \tag{2}$$

However, different methods are used to estimate the mean runoff and the standard deviation per calendar date. Braga *et al.*[14] use the straightforward definition for the arithmetic mean and standard deviation of a discrete set of values on a daily scale as described above. Serinaldi and Kilsby,[15] on the other hand, obtain smoothed estimators by applying a LOESS scatterplot smoothing. This seems to be a tiny detail; we show later that it can have a decisive impact on the resulting spectrum of slopes for the degree distribution.

Here, smoothed estimators for the mean value and the standard deviation are obtained by moving averages given a window length ϱ. For a given moment, i.e., year and day within the year y, i, the averaged mean and standard deviation are obtained from ϱ time series values symmetrically around this moment, i.e., with the reference moment in the center. We consider values between $\varrho = 1$ (the standard equations) and $\varrho = 181$ (roughly a whole year of data is used for smoothing). The parameter ϱ resembles the smoothing parameter of a LOESS estimator. Using this description, the Braga *et al.*[14] deseasonalization procedure simply uses $\varrho = 1$.

The result in all cases is a dimensionless series, where usually a major part of the seasonality is removed, as, e.g., spectral analysis shows. Our point here is to investigate the consequences of the different deseasonalization procedures for the HVG properties, and whether or not these data preprocessing can be recommended when investigating river runoff dynamics with visibility graphs.

C. Tied values

Equation (1) implies that exactly identical values (ties) block the visibility, contrary to values which are smaller by a negligible amount. This fact is not an issue for artificially generated series with a continuous value spectrum; for measured values, it can be an obstacle one has to deal with. In our case, all runoff values provided (consistently given in unit m^3/s) are integers, and at low flow conditions or for small catchment areas, identical values occur frequently. This is, however, an artifact of the accuracy provided or the digitization process of the runoff records; it is safe to assume that for perennial streams, identical values have vanishing probability.

Nonetheless, the HVG algorithm reacts sensibly to the presence of tied values, as will be shown below. The expectation is that removing the ties, e.g., by adding noise of small amplitude or variance, leads to an increase in the mean degree k and in general to increased visibility. An interesting approach to get rid of tied values through imputation, although in the context of ordinal pattern statistics, is provided elsewhere in this volume.[16]

D. Algorithm

It is straightforward to implement the HVG construction into a programming environment. For large time series, issues with memory usage and computation time arise. The first limitation can be overcome through working with sparse matrices, a property that virtually all adjacency matrices deduced from HVGs possess. For this paper, we mainly worked in the R environment but outsourced the proper calculation of the adjacency matrix to precompiled C++ code.

E. Degree distributions

Given the adjacency matrix from a time series using the HVG criterion [Eq. (1)], we determine the number of time series values (nodes) each given node is connected to; this is the degree of that node. Ignoring the very first and last entries of the time series, by construction each node has at least degree $= 2$. It can be shown that for infinitely long, independent random values, the mean degree is $\bar{k} = 4$.[8] The associated degree distribution turns out to be an exponential:

$$P(k) = \frac{1}{3}\left(\frac{2}{3}\right)^{k-2} = \frac{3}{4}e^{-\lambda_c k}, \tag{3}$$

with $\lambda_c = \ln\left(\frac{3}{2}\right) \approx 0.4054$. This result, shown in a lengthy calculation in Ref. 8, is independent from the probability distribution of the time series values, as long as it is the same and there are no (auto-)correlations present. It has been suggested[17] that the degree distribution of both deterministic-chaotic and stochastic but correlated series are as well exponential, with $\lambda < \lambda_c$ for chaotic time series and $\lambda > \lambda_c$ for correlated noise. This would open for the possibility to distinguish between the two process classes by determining the corresponding λ from the HVG degree distribution. However, Ravetti *et al.*[18] provide ample numerical evidence which does not support this suggestion, and the classification based on the criterion seems unfeasible. To the best of our knowledge, there is no analytical expression for the degree distribution of correlated noise available, and it thus remains unclear whether it would be an exponential at all.

Under the hypothesis that the degree distribution is exponential, the mean degree is always between 2 and 4, independent of the type of distribution or the presence of correlations. For periodic time series where no values are repeated within one given period T, the mean degree is

$$\bar{k} = 4\left(1 - \frac{1}{2T}\right)$$

as shown in Ref. 19.

F. Estimation of the slope λ

In practical situations with time series of finite length, even when accepting the hypothesis that the degree distribution is of exponential type, the determination of the slope λ is challenging. The minimum degree is $k = 2$, but in many cases, a peak in the distribution occurs for $k = 3$. The exponential behavior sets in at higher values of k, but where precisely? One needs to define a lower limit for the fit to the exponential function. Similarly, the fit supposedly has to be extended up to a highest k due to small sample effects. In most cases, this cannot be the highest degree actually occurring in the time series, since some of the smaller degrees do not occur at all. A simple suggestion as in Ref. 18, i.e., use $3 \leq k \leq 20$, or $3 \leq k \leq k_{max}$ if it turns out that $k_{max} < 20$, does not seem suitable in general. We will show that different choices for the fit are crucially affecting the λ values obtained. Sensitivity to the choice of the scaling zone (range of k values) has also been observed in Ref. 18 and renders the assumption of exponential decay doubtful at least for real-world applications with limited sample sizes. For Natural (as opposed to Horizontal) Visibility Graphs, correlated noise seems to exhibit power-law behavior instead.[20]

The obvious way to perform the fit to obtain λ estimates is using Ordinary Least Squares (OLS). Here, all degrees occurring have the same weight, implying that high values of k ($k \geq 10$, say, for the time series lengths we are dealing with here), which necessarily come with very small probabilities, can have an important impact on the resulting slope. This renders the estimation procedure unstable, in particular for shorter time series. We consider three alternatives to perform the fit more reliably: weighted least squares (WLS), a fit to the exponential distribution using Maximum Likelihood estimator (MLE), and a fit based on the L moments of the degree distribution (Lmom).

For WLS, we prescribe weights for the probabilities for a given k which are simply the expected $p(k)$ for the random distribution [Eq. (3)]. Both OLS and WLS fits are based on the logarithm of the survival function, which exhibits a straight line with negative slope for a strict exponential. This is the same approach as in Ref. 14. The estimates for MLE and Lmom are obtained by considering the set of k values (in itself a time series) directly, i.e., not using the observed frequencies, and fitting to an exponential distribution. There is the caveat, however, that all degrees are necessarily integers, whereas the estimation method expects a continuous set of values. To analyze whether this mismatch induces biased estimations, we add to each k value a random number uniformly distributed in the interval $[-0.5, 0.5]$, resembling the removal of ties discussed above, although now for the degree distributions. Non-integer degrees also appear when *weighted* HVGs are used, where the difference between the values of linked nodes is used as weights.

So far, no analytical result is known showing that correlated stochastic processes lead to an exponential degree distribution. Such a proof is of course impossible for observed environmental time series as considered here. We therefore fit our degree distributions, and also that for known reference processes, to a flexible class of distributions originating in the Peak Over Threshold approach, the Generalized Pareto Distribution (GPD). GPD has three parameters—location, variance, and shape—and the increased parameter space implies the potential to obtain better fits. This expectation seems to have been confirmed by the comparison provided in Ref. 15.

III. RESULTS

A. HVGs, degree distributions, and the effect of ties for Brazilian river runoff

1. The determination of the slope λ

If we assume an exponential shape for the degree distribution [cf. Eq. (3)], the problem of estimating the slope λ reduces to a linear regression in a semi-logarithmic relationship—between $\ln[p(k)]$ and k. Using the method of least squares, we have to define a lower threshold k_{min}, an upper threshold k_{max}, and a regression method.

Fixing universal k_{min} and k_{max} values for all time series does not seem suitable, as the linear scaling zone apparently depends on the strength of autocorrelations present. For the upper threshold, we rather work with a probability criterion: we stop the regression whenever the probability falls below a critical small value; typical numbers here are e^{-10} to e^{-5}; for the former limit, the corresponding k_{max} ranges from $k_{max} = 16$ for red noise, $k_{max} = 24$ for $\frac{1}{f}$ noise to $k_{max} = 28$ for uncorrelated noise (Fig. 2).

For large values of k, the variance of the probabilities when repeating the analysis is increasing (note, however, the logarithmic vertical scale in Fig. 1). Even for the chosen time series length of 10^8 data points, far exceeding the length of the bulk of observed time series, degrees $k > 50$ practically never occur. The slope of the fits gets steeper when the autocorrelations increase, and in particular for long-range correlated processes such as the power-law (i.e., with a power

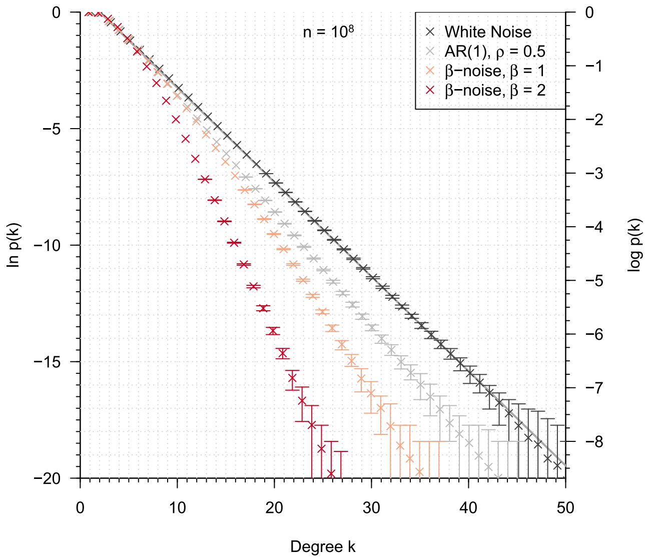

FIG. 2. Degree distributions from long time series of artificial processes. The solid line represents the theoretical exponential for uncorrelated random noise. It is obvious that for 10^8 data points, the white noise shows deviations from the expectation for $k > 40$. For small k, all lines collapse into one but show a curved shape for slightly larger values. Short-ranged [AR(1)] and long-ranged ($1/f$) noise are hard to discern at low to medium k; the values where they really separate are not obtained for short series. The red ($1/f^2$) noise is clearly different from the former two. It is difficult to find a proper scaling zone, at least not of the fixed form [k_{min}, k_{max}] across different processes. Error bars result from 48 repetitions each. For ease of comparison, the logarithms of the probabilities are shown both as natural (left axis) and as decadal (right axis).

spectrum proportional to $1/f^\beta$) noise shown here, the results show a negative curvature for small to medium degrees. For large degrees, it might be difficult to identify whether this curvature flattens into a straight line [as implied by Eq. (3)] for $1/f^\beta$ noise, and slope estimation particularly for short series is difficult. For example, Fig. 3 shows corresponding fits for time series of length $n = 365$ (one year of data at daily resolution) and $n = 10^4$, which is the order of magnitude for the length of many runoff records. For the yearly windows, error bars get large, the white noise degree distribution deviates from the theoretical expectation already for $k > 12$, and for all practical purposes short-ranged [AR(1)] and $1/f$ long-ranged noise become indistinguishable. For $n = 10^4$, the finite size effects are less pronounced, e.g., the onset of deviations from the theoretical curve for white noise is shifted to $k > 23$, and the two noise processes are just about discernable.

As the linear scaling regime for the white noise process stretches out further and further with increasing time series length, followed by a more and more negatively curved part, it can be hypothesized that the curvature is an artifact of the finite length. It is currently unknown how the onset of the curved part, given a certain time series length, differs between processes, impeding the ability to conclude on deviations from the exponential form which are fundamental and not simply related to the finite length of the generated data.

2. Different estimators for the slope of the degree distribution

The plots of Figs. 2 and 3 can be used to determine the slope λ, assuming from here on that the exponential model is a valid description of the observed frequencies of degrees. For its estimation, we use a suite of different methods: OLS and WLS based on the survival function, MLE and L-moments based on the set of degree values directly, and MLE and Lmom supplemented by noise (to get rid of "k-ties") of an amplitude which prevents confusion of k values.

These estimating procedures were applied to time series of length either $n = 365$, $n = 10^4$, or $n = 10^8$. In each case, a number of repetitions were performed, resulting in error bars for the estimated slopes. Figure 4 shows the results. It is unsurprising that the size of the error bars decreases with increasing time series length; for $n = 10^8$ and uncorrelated "white" noise, all methods except MLE estimate a λ value that is very close to the theoretical expectation [Eq. (3)]. However, time series of this length are not available for most practical applications. Hence, for an individual year with daily measurements, i.e., $n = 365$, a reliable estimation of λ is hardly possible [Fig. 4(a)]. Moreover, the different methods for estimation perform differently. For example, MLE has a positive bias independent of process and data length, indicating that the addition of uniform noise should be done when using this method. Adding noise to the degrees before estimating using Lmom also lowers the estimates for λ, although less dramatic. For stronger autocorrelations, the estimates differ more from each other, the error bars are larger, and it becomes unfeasible to determine one "true" expectation value even from artificial data with $n = 10^8$. For the case of red noise, we demonstrate in Fig. 4 that the lowest degree included in the fitting procedure also has an impact on the results, giving higher estimates for the slope for higher cutoff k.

Among the estimators used, the Lmom augmented by uniform noise provides λ values closest to the "asymptotic" values for long time series of uncorrelated noise, but WLS is also a reasonable choice.

3. The impact of ties on the degree distribution

For the Brazilian runoff data, low flow conditions inevitably lead to repeated identical values (ties), not uncommon for runoff data in general. Following the definition for HVG construction [Eq. (1)], tied values are "obstacles": all values "behind" (in both temporal directions) them are invisible. The repeated values are, however, an artifact of the limited measurement resolution; the only situation where exactly identical values would be complete ceasing of the runoff, which is unlikely for most reasonably sized catchments.

The presence of ties leads to shorter degrees than without them and thus also reduces the mean value of the degree distribution. For the runoff data, the connection between $\bar{k}$ and the fraction of ties is extremely strong and linear (Fig. 5). With increasing fraction of ties, $\bar{k}$ runs through practically its whole value spectrum ($2 \leq \bar{k} \leq 4$). Although in our case, ties are merely a nuisance, one has to deal with them.

We, therefore, investigate the consequences of "removing" the ties for the resulting degree distributions. For all tied values, we add either uniformly distributed noise from the interval [$-0.001x_t$, $0.001x_t$] or red noise with an amplitude

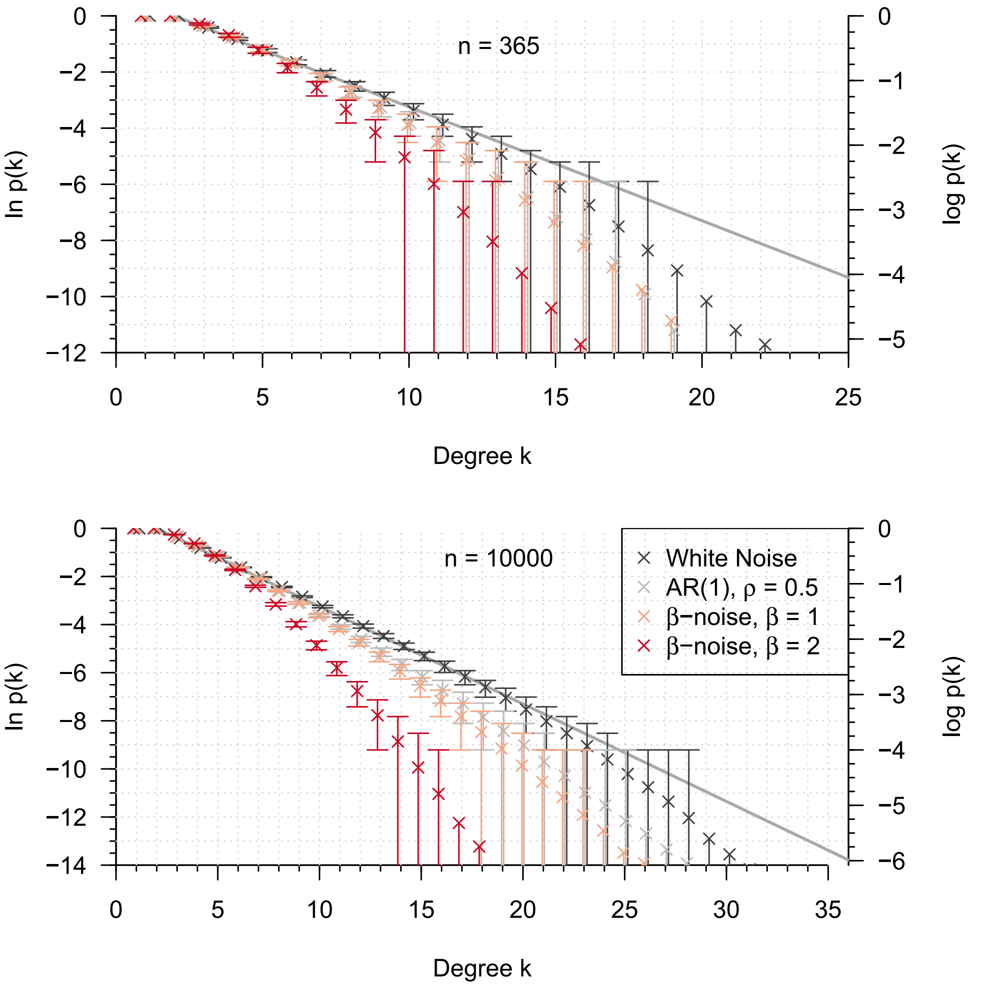

FIG. 3. Similar to Fig. 2, but for smaller time series lengths. Note the different ranges for the degree values.

diminished by the factor 10^4. The resulting time series are visually indistinguishable from the original ones. However, as can be seen from Fig. 5, the consequences for the mean degree are dramatic: the dependence on the number of ties disappears. Another conclusion from Fig. 5 is that for the Brazilian runoff data, $\bar{k}$ is close to 4, the theoretical result for white noise series. Of course, the inverse conclusion that the runoff time series are thus almost perfect noise would be erroneous.

Figure 6 shows that these tiny changes have significant consequences in the case of many ties in the series. Generally, the noise addition moves the degree distribution towards the random case (from below, i.e., from steeper slopes); for around 1% ties, there is hardly any discernible difference, but for higher fractions, the plots begin to diverge, and also the two versions of noise addition are separate from each other, the red noise (or β-noise with $\beta = 2$) being further away from the original distribution. The "correct" type of noise is not obvious to guess. "Long tails" (high values of k) are present only for series with a small amount of ties, resp. after noise has been added to them.

After deseasonalization, the problem has disappeared completely (Fig. 5). The removal of daily means and the standardization to single-day standard deviations leave no room for any repeated values; the ties have disappeared.

4. Consequences of deseasonalization for HVG-derived properties

Although the problem of tied values disappears through the process of deseasonalization, the details of the recipe to deseasonalize induce clear impacts on the resulting degree distribution, a problem which is fully independent from the presence of tied values. This can be seen in Fig. 7. Here, the same three river stations as in Fig. 6 have been used, and two versions for deseasonalization are compared, where only one includes smoothed estimators. One arbitrary chosen year has been picked, but the results are similar for other years as well. The time series graphs are quite similar, although less so for series with a high number of ties.

The important point is the striking difference between the degree distributions for unsmoothed $\varrho = 1$ and smoothed ($\varrho = 91$) deseasonalization. For small degrees, the details of obtaining deseasonalized series are unimportant. However, for larger $k's$, the distributions differ from each other, to the extent that for the conventional estimators, the standard exponential seems to hold with a slightly less steep slope, whereas in the case of smoothed estimators, the tail of the degree distribution stretches to much higher values of k, well above the straight line for the white noise case, possibly implying a better fit would be obtained by fitting to a power law instead. Figure 8 (upper panel) shows this effect for the entire Brazilian river

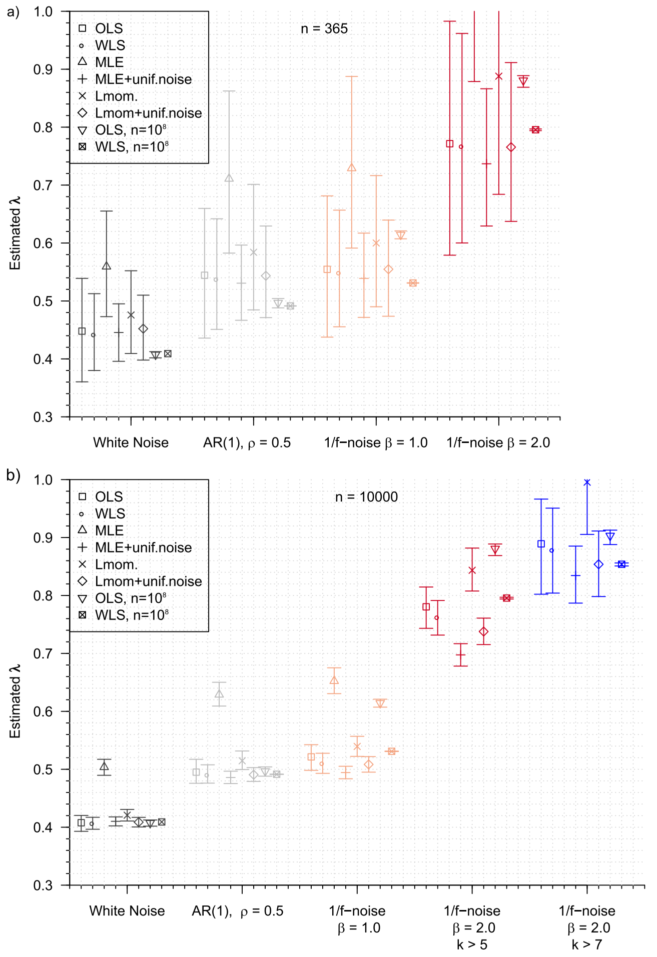

FIG. 4. Estimates of λ for different stochastic processes using various estimators. (a) For $n = 365$ and (b) for $n = 10\,000$. Error bars are 95% confidence intervals for around 50 realizations of the respective processes. For (b), two different values for the lower cutoff are also compared.

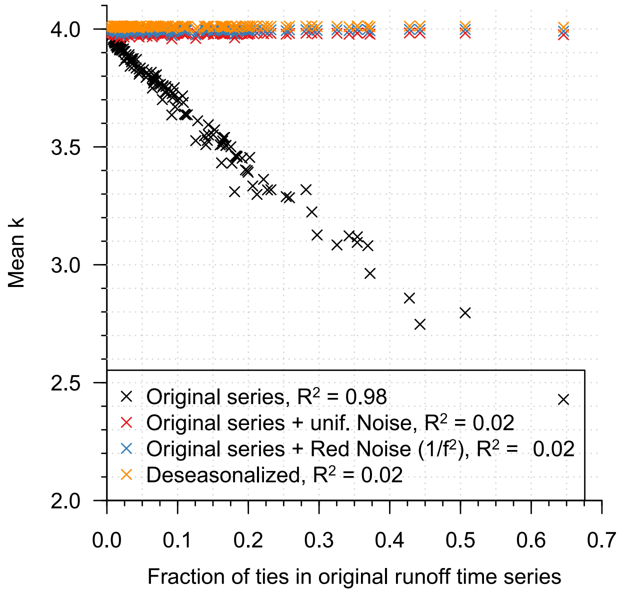

FIG. 5. The mean value of the degree distribution dependent on the fraction of values in the time series which have ties. All runoff time series are shown here, and the fraction varies between well below 1% and more than 60%. Black crosses denote the original time series. If the series are deseasonalized first (green crosses), the dependence disappears completely.

runoff dataset (i.e., HVGs determined for each river individually—for each degree k, the 5th to 95th quantile across rivers is shown): The original time series stretch to high values of k in the degree distribution, i.e., show a "long-tailed" power law like behavior. A removal of the seasonal cycle with $\varrho = 1$ essentially removes this property from the data, while seasonal cycle removal with $\varrho = 91$ appears to largely maintain it. In Fig. S1 (Supplementary material), we show that this effect is equally important for a United States runoff dataset (the HCDN runoff data used in Ref. 15), although rivers in the US runoff dataset show *on average* slightly longer-tailed behavior. It is worth noting that the US data were selected according to minimal management (or most "natural conditions"), while the Brazilian rivers included here are all managed. Nonetheless, the different seasonal cycle removal procedures alone mainly explain the contradicting results obtained in Ref. 14 for Brazilian and in Ref. 15 for United States runoff series.

B. Preprocessing and L moments of the runoff time series

Since the method of L-moments estimation turned out to be suitable for estimating the slopes of degree distributions, we investigate the third and fourth L-moments for the Brazilian river runoff data (Fig. 8, lower panel). The values obtained for these two quantities depend strongly on the preprocessing of the data or their simulation. Simple random shuffling of the time series constrains the L-skewness and L-kurtosis to a narrow region, indicating that the runoff series are far from white noise. However, a standard simulation for runoff series, AR(1) time series where the correlation parameter is empirically estimated from the lag 1 autocorrelation of each series separately, is also confined to a small subset of the original range of values, demonstrating that the AR(1) model derived from the time series does not retain the key characteristics of the time series and is thus not suitable for these data. Small amounts of noise are not deleterious for the L-moments' distributions, but deseasonalization, and in particular the direct unsmoothed estimation of daily mean and standard deviation, has a strong effect. Runoff dynamics might be seen as a superposition of short-term correlations with "long-tailed" highly visible nodes, often flood events.

C. GPD fits to degree distributions

As a flexible alternative to the exponential distribution, degree distributions might also be fitted to Generalized Pareto Distributions (GPDs), advocated in Ref. 15 as a superior method in a Maximum Likelihood sense. Long-tailed GPDs have a shape parameter $\xi > 0$. Parameter estimation can be conveniently performed using L-moments again.

Figure 9 compares the ξ estimates for the river runoff data in original form, and the same preprocessing or simulation versions just discussed. All shape parameters obtained are positive, and the addition of small amounts of noise does not change much. When shuffling the data or producing AR(1) versions of them, the ξ values get close to zero, and are clearly different from the original ones. Deseasonalization lowers the estimates, in particular in the unsmoothed case; however, once we foresee a smoothing for mean and standard deviation, the

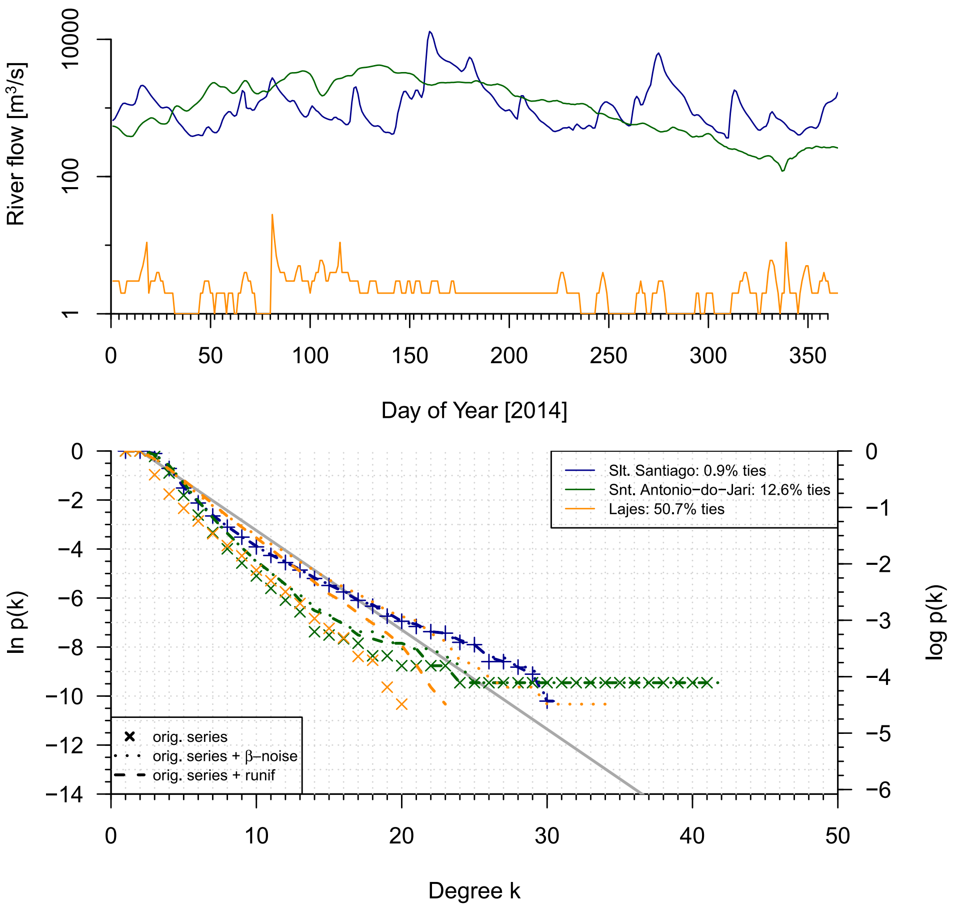

FIG. 6. Attempts to remove ties by adding noise to the time series. Upper panel: three selected time series with low, intermediate, and high fraction of ties (note the logarithmic scale), illustrating the origin of the problem—low flow values. Lower panel: the resulting degree distributions for the original series, for series contaminated by uniform noise, and by β-noise ($\beta = 2$).

window length for the moving average is unimportant. Thus, the shape parameter of L-moments based GPD fits to the degree distributions is a robust property of the runoff records.

D. Connection between the slope of the degree distribution and mean runoff

We have seen that in virtually all cases, the values for λ obtained with the different methods are larger (the degree distribution is steeper) than for the random case. But what determines the value for λ for an individual station? One approach towards this direction is the scatter diagram shown in Fig. 10. By and large, the catchment size determines the long-term mean runoff (in absolute units, such as m^3/s), so the horizontal axis of Fig. 10 essentially represents catchment area. When estimating the λ values on the vertical axis through simple OLS, no sensible connection is apparent. However, utilizing WLS instead with a lower cutoff of $k = 5$, the scatterplot reveals that the slopes are getting steeper with increasing mean runoff or catchment size. Correlations in the runoff record thus have the tendency of getting stronger for larger catchments. This is in accordance with a common understanding of hydrological processes, as runoff from larger rivers integrates over a larger catchment and thus an extended river network and over a wide range of transit times, where, e.g., localized individual precipitation events are smoothed out and are less important.

E. The effect of dam construction on the slopes λ

A special property of the runoff data from Brazil used here is that each and every station is currently related to water power generation. The data are maintained by the ONS, the federal institution which is in charge of the water power system in Brazil. The system of water power generation is expanded since some decades, and the dataset used only contains stations where a dam was constructed during the time of the record. Although great care and effort was devoted to the construction of so-called natural discharge (the sum of observed and consumed water),[21] we analyzed the extent of impact of dam construction on the runoff series in a simple manner: for each station where sufficient data were available, a period of 20 years prior to the construction and 20 years after it was identified [we observe that the obtained result is very similar if 12-year periods are used, cf. Fig. S2 (Supplementary material)]. A buffer time of 3 years on each side of the dam construction year was introduced, and from the remaining 17 years each, the slopes for the degree distribution were calculated. The procedure left a total of 38 stations. As Fig. 11 shows, the clear majority of stations exhibits an increase in the λ value across the dam construction year; stronger correlations in the HVG prevail after completion of the dam. This result is in principle a confirmation of that of Ref. 14, which also finds significant positive (linear) trends in λ; however, when calculating the λ's from

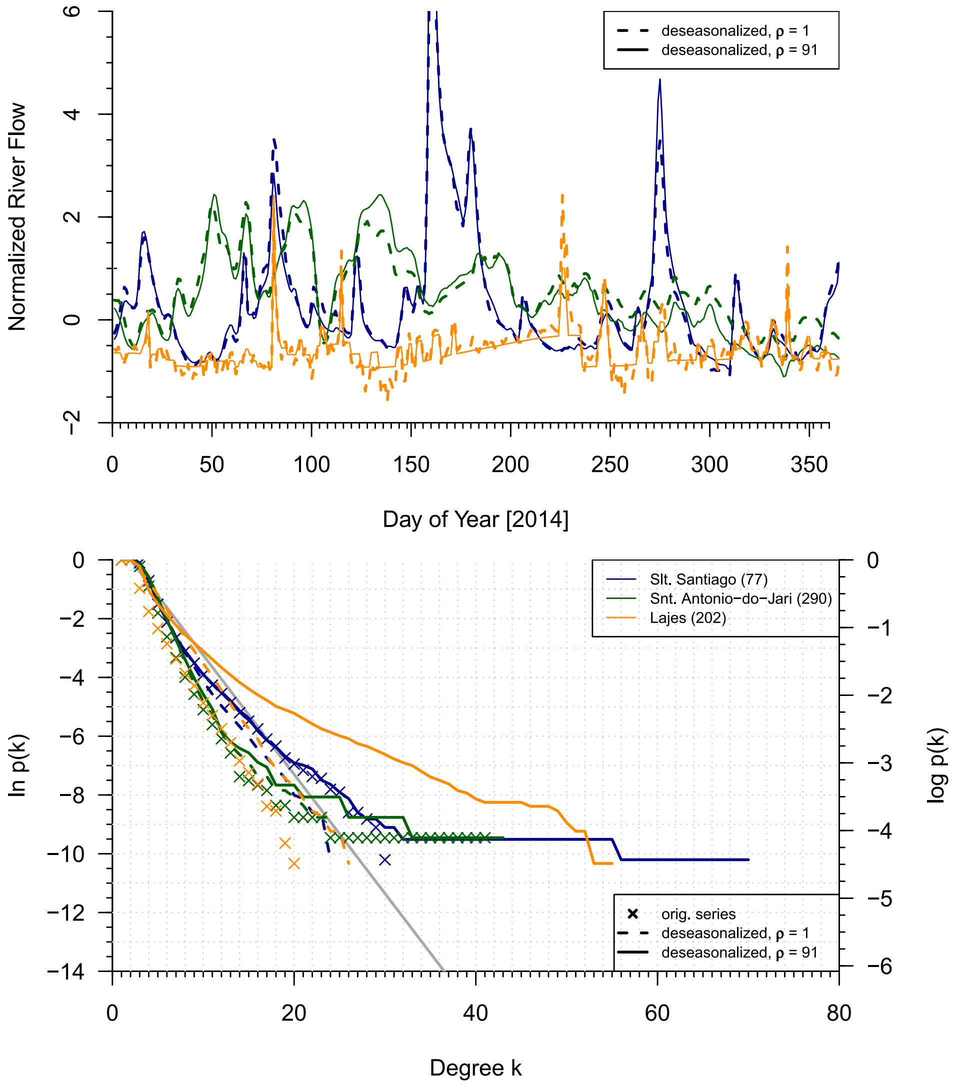

FIG. 7. Effect of different deseasonalization procedures on the degree distribution for three randomly selected rivers in an exemplary year (2014) in (a) and for the degree distribution derived from the whole time series (b).

individual years prior to the dam construction, not more than roughly half of them show an increase, and some changed positively only in the after-dam construction period. This indicates that the observed positive change is indeed due to the dam construction.

IV. SUMMARY AND CONCLUSIONS

We have investigated the method of Horizontal Visibility Graphs (HVGs) in the context of a set of Brazilian river runoff data, focusing on the degree distributions derived from the networks. We considered the known exponential distribution for uncorrelated white noise as a benchmark for the measured time series. The estimation of the slope of the logarithmized distribution against degrees turned out to be surprisingly delicate, and required considerable care:

- Different deseasonalization procedures induce different degrees of sub-exponential decay in the degree distributions; it is likely that this explains the apparently contradictory results of long-tailed behavior of North American[15] vs. short-tailed behavior of Brazilian[14] rivers.
- Real-world time series are typically short; estimation procedures for HVG-based metrics such as the slope λ in the degree distribution have confidence intervals that can be broad, and some systematic biases; thus, care is needed in interpreting different estimators.
- Ties in the data affect the mean degree obtained by HVG analysis.

For the case of the Brazilian runoff dataset, we show that

- WLS estimates mostly quantify short-term dynamics (i.e., for relatively small k values) and are thus related to the (short-term) noise behavior. The latter relates to general catchment characteristics, the degree of management or climatology. In contrast, metrics related to the tail of the k-distribution (L-skewness, L-kurtosis, shape parameter of the GPD) are mostly related to "high visibility events," which appear to be largely independent from short-term noise.

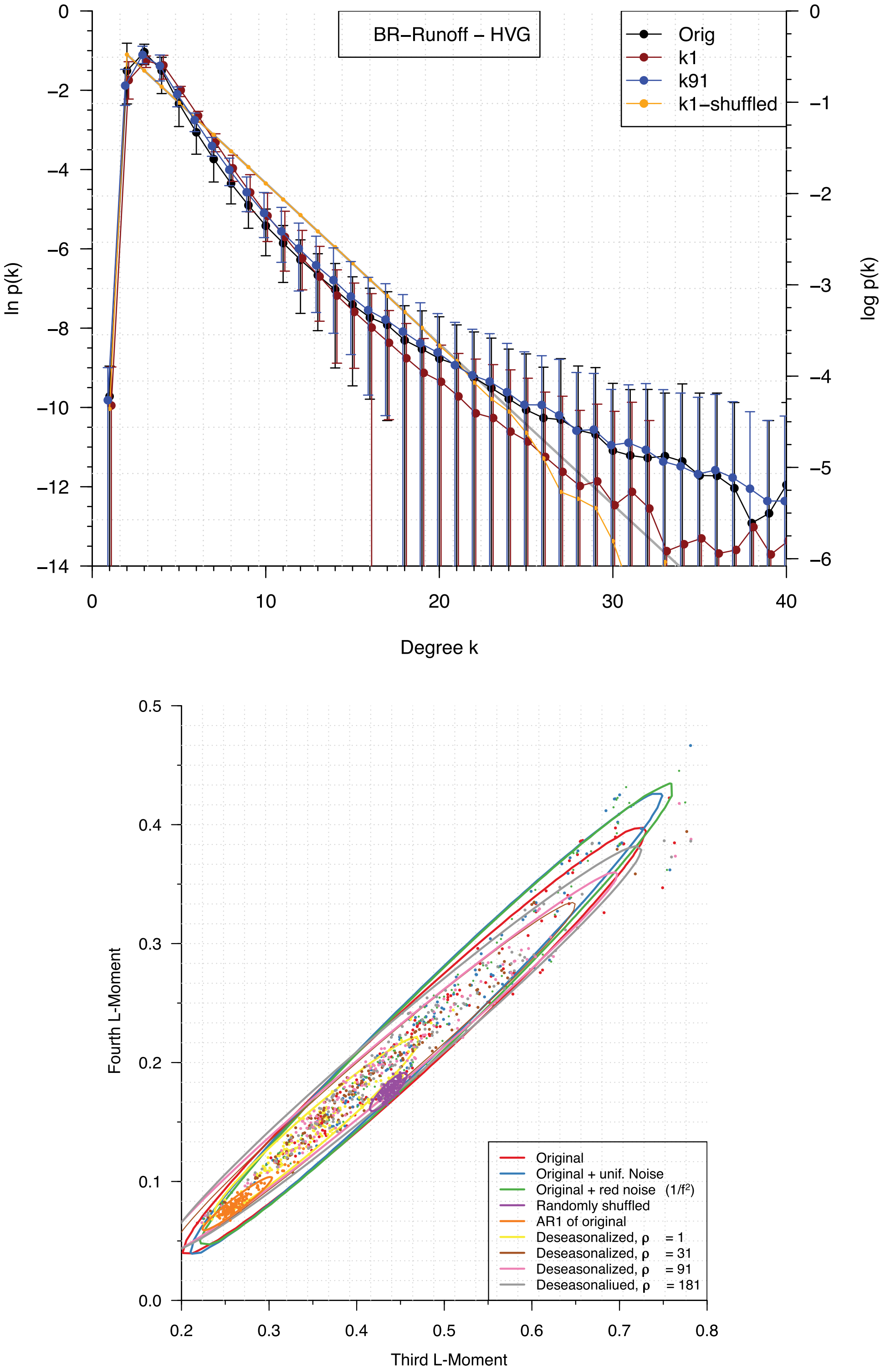

FIG. 8. Systematic effect of different deseasonalization and data preprocessing procedures on (top) degree distributions and (bottom) L-moments estimated from the entire Brazilian river runoff dataset. It becomes clear that the dynamics of the original time series differ substantially from correlated or uncorrelated noise ("AR1 of Original" and "Randomly shuffled," respectively) and that conventional ($\varrho = 1$) deseasonalization induces dynamics that is closer to correlated noise than the original time series or those that are deseasonalized with a smoothed seasonal cycle ($\varrho > 1$). Ellipses are 95% areas of respective fits to two-dimensional Gaussian distributions.

- Estimates of λ for Brazilian runoff show indeed changes in their short-term characteristics, and these trends are related to the construction of dams.

On a more general note, HVGs are a useful tool, but sensitive to short, noisy time series or data preprocessing or transformations of any kind (here, deseasonalization). An alternative might be weighted HVGs, as used in Ref. 10, for instance, where the amplitude information is retained, and the time series can be fully reconstructed from the HVG network.

SUPPLEMENTARY MATERIAL

See Supplementary material for two additional figures. The first figure is similar to the upper panel of Fig. 8, but the runoff time series are taken from the Hydro Climatic Data Network (HCDN) maintained by the US Geological Survey, identical to the dataset used by Ref. 15, where the selection was based on minimal human activity in the respective basins. The second figure is similar to Fig. 11, but here we used a window of 12 years before and after the dam construction,

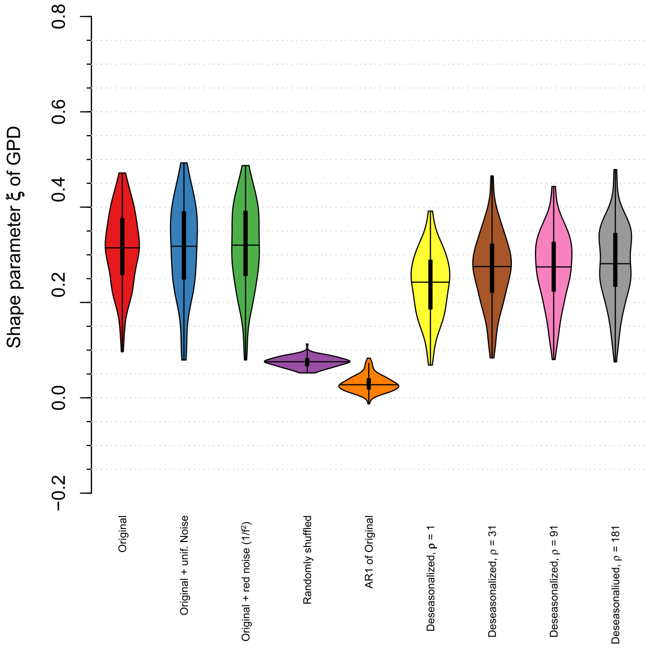

FIG. 9. "Shape" parameter ξ of the Generalized Pareto Distribution estimated from the degree distribution via L-moments. Almost all Brazilian river runoff records show "long-tailed" behavior in their degree distributions (positive ξ values). The violin plots contain estimates from all runoff records in our set.

increasing the number of gauges in the analysis, without changing the main conclusion.

ACKNOWLEDGMENTS

Discussions with F. Serinaldi are gratefully acknowledged. One of the authors (H.L.) wants to thank the people at the Physics Department at the Universidad Federal de Alagoas (UFAL) in Maceió, Brazil, for the warm hospitality during the sabbatical stay where this work was initiated.

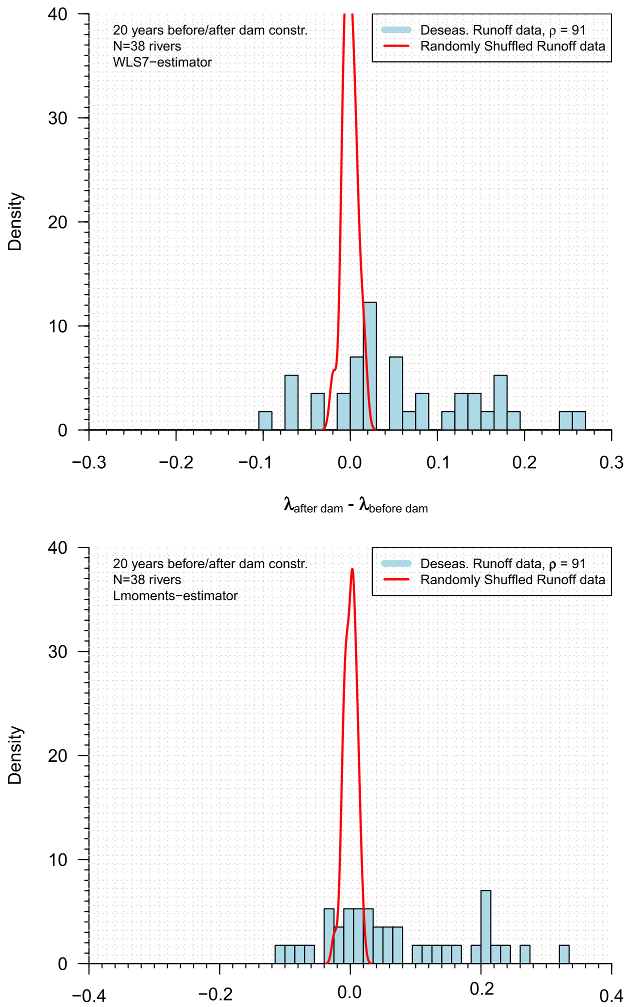

FIG. 11. Slope estimates before and after dam construction at 38 of the river stations.

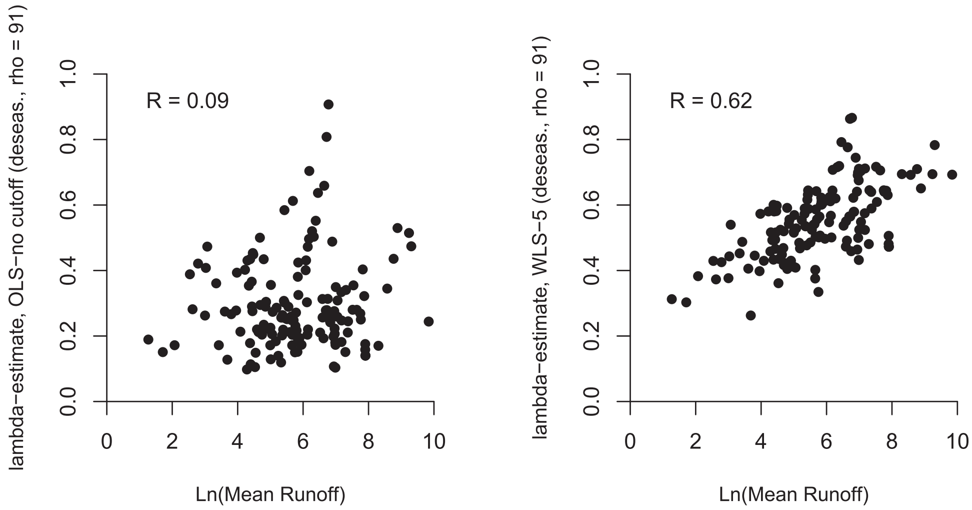

FIG. 10. Scatter diagrams of the logarithm of the mean runoff for all Brazilian river stations and λ estimated either with OLS (left panel) or with WLS (right panel).

[1] S. H. Strogatz, Nature **410**, 268–276 (2001).

[2] R. V. Solé and J. M. Montoya, Proc. R. Soc. Biol. Sci. **268**(1480), 2039–2045 (2001).

[3] R. Albert and A. L. Barabasi, Rev. Mod. Phys. **74**(1), 47–97 (2002).

[4] J. D. Phillips, W. Schwanghart, and T. Heckmann, Earth Sci. Rev. **143**, 147–160 (2015).

[5] N. Attar and S. Aliakbary, Chaos **27**(9), 091102 (2017).

[6] M. Barthélémy and L. A. N. Amaral, Phys. Rev. Lett. **82**, 3180–3183 (1999).

[7] A.-L. Barabási, R. Albert, and H. Jeong, Physica A **281**, 69–77 (2000).

[8] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, Phys. Rev. E **80**(4), 046103 (2009).

[9] L. Lacasa, B. Luque, and A. Nuñez, in *New Frontiers in Graph Theory*, edited by Y. Zhang (InTech, 2012).

[10] B. A. Gonçalves, L. Carpi, O. A. Rosso, and M. G. Ravetti, Physica A **464**, 93–102 (2016).

[11] H. E. Hurst, "Long-term storage capacity of reservoirs," Trans. Am. Soc. Civil Eng. **116**, 770–799 (1951).

[12] N. W. Arnell and S. N. Gosling, Clim. Change **134**(3), 387–401 (2016).

[13] L. Gudmundsson, S. I. Seneviratne, and X. Zhang, Nat. Clim. Change **7**, 813 (2017).

[14] A. C. Braga, L. G. A. Alves, L. S. Costa, A. A. Ribeiro, M. M. A. de Jesus, A. A. Tateishi, and H. V. Ribeiro, Physica A **444**, 1003–1011 (2016).

[15] F. Serinaldi and C. G. Kilsby, Physica A **450**, 585–600 (2016).

[16] F. Traversaro, F. O. Redelico, M. R. Risk, A. C. Frery, and O. A. Rosso, Chaos **28**, 075502 (2018).

[17] L. Lacasa and R. Toral, Phys. Rev. E **82**(3), 036120 (2010).

[18] M. G. Ravetti, L. C. Carpi, B. A. Gonçalves, A. C. Frery, and O. A. Rosso, PLoS ONE **9**(9), e108004 (2014).

[19] A. Nuñez, L. Lacasa, E. Valero, J. P. Gómez, and B. Luque, Int. J. Bifurcat. Chaos **22**(07), 1250160 (2012).

[20] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuño, Proc. Natl Acad. Sci. **105**(13), 4972–4975 (2008).

[21] F. A. Oliveira, E. L. de Melo, J. C. Figueiredo, F. F. Pruski, and R. d. G. Rodriguez, Rev. Bras. Recur. Hídricos **13**(3), 191–197 (2007).